A TWO-SAMPLE SEQUENTIAL RANK TEST BY SEN INVESTIGATED BY STOCHASTIC SIMULATION

Grete U. Fenstad and Eva Skovlund
Institute of Mathematics and Department of Informatics
University of Oslo, Norway

Correspondence:

Eva Skovlund
Department of Informatics
P O Box 1080 Blindern
0316 Oslo 3
Norway

Abstract

We have investigated the properties of a two-sample sequential rank test proposed by Sen (1981) by means of stochastic simulation. The test is claimed to be non-parametric. Our simulations show that the significance level and power are not far from the respective nominal values for most of the selected distributions. However, the simulated values deviate significantly from the nominal values for some distributions. Compared to another two-sample sequential rank test proposed by Skovlund and Walløe (1988), Sen's test reaches a decision somewhat earlier, but the significance level and power are more dependent on the shape of the distributions of observations, and the test statistic is much more cumbersome to calculate.

1. INTRODUCTION

Skovlund and Walløe (1988) have recently developed a sequential two-sample rank test by means of stochastic simulation. The test is based on the Wilcoxon-Mann-Whitney two-sample test for fixed sample sizes and has been shown to be robust and approximately distribution free. It has been suggested to the authors that possibly a better distribution free sequential two-sample test could be developed from the Type D test suggested by Sen (1981, p. 255). The present paper describes the development and exploration of such a test.

2. THEORY

Following Sen (1981; pp 255, 264) let $\{X_i, i \geq 1\}$ be a sequence of independent observations and we assume that

$$X_i = \Delta c_i + \varepsilon_i, \quad i \geq 1$$

where the c_i are known constants, 0 or 1, and the ε_i are independent identically distributed random variables with density $f(t)$. The problem is to test $H_0 : \Delta = 0$ versus $H_1 : \Delta = \Delta_1 (> 0)$. In the frame of sequential clinical trials the X_i are the responses of the patients either in the treatment group ($c_i = 1$) or in the control group ($c_i = 0$). Under H_0 the responses in both groups have the same distribution, i.e. the treatment has no effect.

We observe the X_i sequentially. Let m be the number of observations in the treatment group among the n first observations and let W_n be Wilcoxon's rank-sum statistic. The statistic

$$U_n = (W_n - m(n+1)/2)/(\sqrt{m(n-m)}(n+1))$$

has mean 0 and variance 1/12(n+1) under H_0 and under alternative Δ_1 close to 0 the asymptotic mean of U_n is $\Delta_1 \gamma(f)$ if $m/n \to p$ where

$$\gamma(f) = \sqrt{p(1-p)} \cdot \int_{-\infty}^{+\infty} f^2(t)dt = \phi(p) \cdot \kappa(f)$$

If b is subtracted from all the observations in the treatment group before W_n is calculated, we denote the new statistics $W_n(b)$ and $U_n(b)$ respectively.

Sen suggests basing a sequential test on

$$Z_n = \Delta_1 D_n U_n(\frac{1}{2}\Delta_1) 12(n+1)$$

where $\{D_n\}$ is a consistent sequence of estimators of $\gamma(f)$. (If $\gamma(f)$ is known, $\gamma(f)$ may replace D_n in Z_n.) Starting with an initial sample of size n_0, proceed as usual in sequential probability ratio tests until one of the inequalities

$$\ln \frac{\beta}{1-\alpha} = b < Z_n < a = \ln \frac{1-\beta}{\alpha}$$

is violated. Let M be the first n such that $Z_M \notin \langle b, a \rangle$, then accept H_0 if $Z_M \leq b$ and accept H_1 if $Z_M \geq a$. Here α is the nominal significance level and $1-\beta$ is the nominal power.

We now return to the construction of a consistent sequence of estimators of $\gamma(f)$. A consistent sequence of estimators of $\phi(p) = \sqrt{p(1-p)}$ is obviously $\phi(\frac{m}{n}) = \sqrt{\frac{m}{n}(1-\frac{m}{n})}$, and consistent sequences of estimators of the integral $\kappa(f) = \int_{-\infty}^{+\infty} f^2(t)dt$ are found on p. 264 in Sen. In our case we base our sequence on Wilcoxon's rank-sum statistic and utilize the close connection between this statistic and all differences between the observations in the treatment group and in the control group. Let $D_{(1)} < D_{(2)} < \cdots < D_{(m(n-m))}$ be these differences ordered, then (see Lehmann (1975; Theorem 4, p. 87))

$$D_{(l)} \leq \Delta \iff W_n(\Delta) \leq m(n-m) + \frac{1}{2}m(m+1) - l.$$

This leads after some calculation to the following estimator of $\kappa(f)$

$$K_n = \frac{2u_{\epsilon/2}}{D_{(l_2)} - D_{(l_1)}} \sqrt{\frac{n^2-1}{12} \cdot \frac{1}{n\frac{m}{n}(1-\frac{m}{n})} \frac{1}{n+1}}$$

where

$$\left.\begin{matrix} l_2 - 1 \\ l_1 \end{matrix}\right\} = \frac{m(n-m)}{2} \pm u_{\epsilon/2} \sqrt{\frac{n^2-1}{12} n \frac{m}{n}(1-\frac{m}{n})}$$

and $u_{\epsilon/2}$ is the upper $\frac{\epsilon}{2}$-fractile of the $N(0,1)$-distribution.

This has all been developed under the assumption that the c_i are fixed. However, we consider the c_i as independent identically random variables with $P(c_i = 1) = p = 1 - P(c_i = 0)$, and we examine the behaviour of Sen's test in case $p = 1/2$.

3. SIMULATION

To examine the properties of Sen's test, we have used computer simulation. The simulation programs were written in the programming language SIMULA (Birtwistle et al, 1983). The pseudo-random number generator in SIMULA is a multiplicative congruential generator (Bratley et al, 1983). The simulations were performed on a DEC 2060/2065 computer at the University of Oslo.

We are especially interested in modelling sequential clinical trials, and have therefore restricted our simulations to differences in treatment effect which are usually clinically relevant. The simulations have therefore been performed for differences ranging from $\Delta = 0.5$ to $\Delta = 2$ in distributions with standard deviation 1. In the simulation model the response for each patient is supposed to be known before a new patient is included. Each included patient is randomized either to a treatment group or to a control group with probability $p = 1/2$. After randomization a response is drawn from a given distribution. Under the null hypothesis (no treatment difference) the responses are drawn from the same distribution. Under the alternative hypothesis the responses in the two groups are drawn from similar distributions with equal variances but different expectations.

Based on the responses of the patients included in the trial so far (at least one in each group), the test statistic is calculated. If neither of the boundaries is crossed, a new patient is included, and the test statistic is again calculated. When one of the boundaries is crossed, the result (acceptance or rejection of the null hypothesis) and the number of patients used is registered. One such sequence is repeated N times.

In a non-parametric situation the value of $\kappa(f) = \int f^2(t)dt$ has to be estimated, but if the distribution is known, the exact value may be used. For the normal, uniform, logistic and double exponential distributions we have calculated $\kappa(f)$ when $\sigma^2 = 1$. The alternative Δ is expressed in units of the standard deviation. The Cauchy distribution $f(x) = r/(\pi(r^2 + x^2))$ has no variance. To obtain the same probability between -1 and +1 as for the standard normal distribution, the parameter $r = 0.54427$ is chosen.

The estimation of $\kappa(f)$ is described at the end of the previous section. The estimation starts when a chosen number of patients n_0 has been included in the trial. We have used $\epsilon = 0.05$.

4. RESULTS

In the test statistic Z_n we use either $D_n = \phi(\frac{m}{n})\kappa(f)$ or $D_n = \phi(\frac{m}{n})K_n$. In the first case the calculation of Z_n is started when at least one patient is included in each group. In the second case the calculation starts when n_0 patients are included.

First we have examined the properties of the test in the situation where the distribution is known. Here $\kappa(f)$ was calculated for the normal, uniform, double exponential, logistic and Cauchy distributions. Hence, this version of the test is not supposed to be distribution free. In Table 1, the simulated significance level and power of Sen's test are shown together with $\kappa(f)$ for each chosen type of distribution. The nominal values of the significance level and power are $\alpha = 0.05$ and $1 - \beta = 0.95$. The alternative hypothesis is $\Delta = 1$. Each result is based on $N = 10000$ simulations. The results are close to the nominal values except for the uniform and the Cauchy distribution where the significance level is a little too large and the power a little too small. Similar results were obtained for other choices of α and $1 - \beta$.

We next examined how the choice of number of initial observations n_0 influenced the significance level and the power when $\kappa(f)$ was estimated. Table 2 shows this as well as the mean and median number of patients included before the trial is stopped, and the mean and median of $D_n = \phi(\frac{m}{n}) \cdot K_n$. The responses are drawn from normal distributions, and the alternative is $\Delta = 1$. For small values of n_0 the simulated significance level and power deviate substantially from the nominal values 0.05 and 0.95. The estimates D_n also deviate from the calculated value $\gamma(f) = 1/(4\sqrt{\pi}) = 0.141$. Only when $n_0 = 10$ or $n_0 = 12$, are the results satisfactory. For larger values of n_0, the test becomes too conservative, as the number of patients included becomes larger than actually necessary. For values of n_0 of about 30, the procedure will no longer be sequential, because the trial is stopped when exactly n_0 patients are included. Based on the results shown in Table 2, we have chosen to use $n_0 = 10$ for further investigation of the properties of the test.

We have also compared the simulation results from the two types of situations (i) distribution known and thus $\kappa(f)$ calculated and (ii) distribution unknown and thus $\kappa(f)$ estimated by K_n. Such comparisons are shown in Tables 3 and 4. In Table 3, the significance level and power of the test using $\kappa(f)$ and K_n are

compared when the responses are drawn from normal distributions where Δ is the alternative. The last column of Table 3 gives $D_n = \phi(\frac{m}{n}) \cdot K_n$ which is the estimates of the exact value $\gamma(f) = 1/(4\sqrt{\pi}) = 0.141$. For small values of Δ, the simulated significance level and power and D_n are satisfactory. For larger values of Δ, the test becomes too conservative, especially when $\kappa(f)$ is estimated.

Table 4 shows the robustness properties of the test. In the first part of the table, $\kappa(f)$ is calculated for the normal distribution even when the responses come from other distributions, e.g. contaminated normal or skew distributions. The nominal significance level and power are again $\alpha = 0.05$ and $1 - \beta = 0.95$, and $\Delta = 1$. The test seems to be quite robust, but not distribution free. The other part of the table shows the results of the test when $\kappa(f)$ is estimated by K_n. As far as contaminated normal distributions are concerned, this procedure is more robust, but even if the results are slightly improved also for the other distributions, they still deviate quite substantially from the nominal values.

5. DISCUSSION

The development of Sen's Type D test assumes among other things small values of the treatment difference Δ. However, in clinical trials the discovery of very small treatment differences is often not of interest. Trials including large numbers of patients are also difficult to handle. We have therefore restricted our simulations to $\Delta \geq 0.5$.

The test has been developed under the assumption that the c_i are fixed. Our simulations have been performed both with fixed c_i and with c_i as random variables where $P(c_i = 1) = P(c_i = 0) = 1/2$. The results were more or less identical, and we have chosen to present the results of random c_i which is the more common situation in clinical trials.

When calculating or estimating the value of $\gamma(f)$, we have presented the results of simulations based on $\gamma(f) = \phi(\frac{m}{n}) \cdot \kappa(f)$ where m is the number of patients included in the treatment group and n is the total number of patients included. Another possibility would have been to put $\phi(p) = 1/2$ instead of the estimate $\phi(\frac{m}{n})$. Then $\gamma(f) = \frac{1}{2}\kappa(f)$. As $m/n \to 1/2$ throughout the trial, there is hardly any difference between using $\phi(p)$ and $\phi(\frac{m}{n})$ for small values of Δ. When Δ is large, however, only few patients are included, and m/n will often deviate substantially from $1/2$. A small simulation study comparing the two alternatives has confirmed that the results are slightly better when using $\phi(\frac{m}{n})$ than when using $\phi(p) = 1/2$.

The number of patients n_0 included before starting the estimation of $\gamma(f)$ is of importance to the test result. If a small n_0 is chosen, the estimates deviate substantially from the theoretical values, and the significance level and power of the test become too large and too small, respectively. For large values of n_0 the test becomes too conservative. We have found the value $n_0 = 10$ to be satisfactory for our range of Δ. The estimates of $\gamma(f)$ are then close to the theoretical values, and the significance level and power are close to the respective nominal values when $\Delta \leq 1.25$. The reason that the test becomes too conservative when $\Delta > 1.25$ is that the choice of $n_0 = 10$ results in inclusion of more patients than would actually have been necessary to reach a conclusion if $\kappa(f)$ had been calculated. The choice of a smaller n_0 could reduce this problem, but then the estimate K_n would deviate even more from $\kappa(f)$ and the test would not be exact anyway. Hence, $\kappa(f)$ should probably not be estimated when $\Delta > 1.25$.

For the normal distribution $N(0,1)$, $\kappa(f) = 1/(2\sqrt{\pi})$. Using this value even when

the observations are not normal, shows that this version of the test is actually quite robust. When $\kappa(f)$ is estimated instead, the test is slightly more robust, especially when the observations come from contaminated normal distributions. The difference is slight, however, and even when the estimation method is used, the test is dependent on the shape of the distributions.

In addition to investigating the properties of the type D test, we have examined the type C test (Sen, 1981, p. 254). The properties of the type C test have been examined in the same manner as the type D test. A comparison of the two tests shows that the type C test is even more conservative than the type D test, it is less robust and definitely not distribution free. Even when the exact value of $\kappa(f)$ is calculated, the real values of the significance level and power deviate substantially from the nominal values. Undoubtedly the type D test is a better alternative than the type C test.

Sen's type D test has been investigated as alternative to the test by Skovlund and Walløe (1988). One of the advantages of Sen's test is that it includes a smaller number of patients to reach a conclusion. If the alternative is $\Delta = 1$ and the observations are $N(0,1)$ under H_0, Skovlund and Walløe's test needs 22 patients (median) under H_0 and 29 under H_1, while Sen's test needs 22 patients both under H_0 and under H_1. The robustness properties of the two tests are not very different; for both tests the simulated power is close to the nominal value over a range of distributions. There is however a tendency towards Sen's test being a little more robust as far as contaminated normal distributions are concerned. On the other hand the simulated significance level is much closer to the nominal value for Skovlund and Walløe's test than for Sen's test where it tends to be too small, even when $\kappa(f)$ is estimated. Contrary to the test by Skovlund and Walløe, Sen's test seems not to be distribution free. If $\kappa(f)$ is estimated, Sen's test is also much more cumbersome to use. Despite the theoretical basis of Sen's test, the test by Skovlund and Walløe therefore seems to be more useful for most practical situations when $\Delta \geq 0.5$.

Acknowledgements

We wish to thank the anonymous referee in *JSCS* for suggesting the investigation of Sen's test. We also wish to thank Lars Walløe for valuable discussions and comments on the manuscript. E.S. was financially supported by The Norwegian Research Council for Science and the Humanities.

REFERENCES.

Birtwistle, G.M., Dahl, O.J., Myhrhaug, B. and Nygaard, K. (1983). *Simula Begin.* Chartwell-Bratt Ltd, Bromley.

Bratley, P., Fox, B.L. and Schrage, L.E. (1983). *A Guide to Simulation.* Springer-Verlag, New York.

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, San Francisco.

Sen, P.K. (1981). *Sequential Nonparametrics: Invariance Principles and Statistical Inference.* Wiley, New York.

Skovlund, E. and Walløe, L. (1988). A Sequential Two-Sample Test Developed by Stochastic Simulation, *Journal of Statistical Computation and Simulation*, **29**, 87-104.

Distribution	$\kappa(f)$	$\hat{\alpha}$	$1-\hat{\beta}$
Normal (0,1)	$1/(2\sqrt{\pi})$	0.0453	0.9503
Uniform $(0,\sqrt{12})$	$1/\sqrt{12}$	0.0610	0.9395
Double exp $(\lambda=\sqrt{2})$	$\sqrt{2}/4$	0.0481	0.9497
Logistic $(b=\sqrt{3}/\pi)$	$\pi/(6\sqrt{3})$	0.0466	0.9527
Cauchy $(r=0.54427)$	$1/(6r)$	0.0600	0.9426

Table 1.

Simulated values of the significance level and power when the distribution is known and thus the exact value of $\kappa(f)$ can be calculated. The nominal significance level is 0.05 and the nominal power is 0.95. The alternative hypothesis is $\Delta = 1$. Behind each result lie $N = 10000$ simulations.

n_0	$\hat{\alpha}$	# pats. mean median	D_n mean median	$1-\hat{\beta}$	# pats. mean median	D_n mean median
3	0.165	16.4	1.623	0.831	16.9	1.194
		10.0	0.172		11.0	0.170
5	0.076	23.3	0.174	0.901	23.8	0.177
		19.0	0.139		19.0	0.140
8	0.070	25.0	0.144	0.924	26.8	0.144
		21.0	0.136		21.0	0.136
10	0.056	25.7	0.142	0.925	26.9	0.142
		21.0	0.135		22.0	0.134
12	0.051	26.4	0.141	0.938	27.9	0.139
		22.0	0.135		22.0	0.133
15	0.043	28.1	0.139	0.945	29.3	0.139
		22.0	0.135		23.0	0.133
18	0.037	29.4	0.138	0.956	31.1	0.136
		23.0	0.133		24.0	0.132
20	0.037	30.0	0.137	0.959	32.3	0.136
		24.0	0.133		26.0	0.132

Table 2

Simulated values of the significance level and power for different values of n_0. The nominal values are $\alpha = 0.05$ and $1-\beta = 0.95$. Here n_0 is the number of patients included before the estimation of $\gamma(f)$ is started. The mean and median number of patients included and the mean and median of D_n are also shown. The theoretical value of $\gamma(f)$ is 0.141. Behind each result lie $N = 1000$ simulations. The responses are drawn from normal distributions with $\sigma = 1$, and the alternative is $\Delta = 1$.

	$\kappa(f)$				K_n				
Δ	$\hat{\alpha}$	# pats mean median	$1-\hat{\beta}$	# pats mean median	$\hat{\alpha}$	# pats mean median	$1-\hat{\beta}$	# pats mean median	D_n mean median
0.5	0.0488	93.4 74.0	0.9539	94.8 75.0	0.068	90.4 76.0	0.952	92.8 75.0	0.145 0.140
0.75	0.0446	43.2 35.0	0.9501	43.1 35.0	0.0612	43.1 35.0	0.9388	42.9 34.0	0.145 0.139
1.0	0.0453	25.3 20.0	0.9503	25.5 20.0	0.0506	26.7 22.0	0.9429	26.4 22.0	0.141 0.135
1.25	0.0463	17.0 14.0	0.9508	17.0 14.0	0.0454	19.4 16.0	0.9558	19.5 16.0	0.138 0.130
1.5	0.0507	12.3 10.0	0.9506	12.2 10.0	0.0365	15.6 13.0	0.9652	15.7 13.0	0.133 0.125
1.75	0.0498	9.4 8.0	0.9508	9.4 8.0	0.0268	13.5 11.0	0.9723	13.5 11.0	0.130 0.121
2.0	0.0401	8.2 8.0	0.9631	8.2 7.0	0.0203	12.2 10.0	0.9817	12.3 10.0	0.127 0.119

Table 3

Simulated values of the significance level and power for different values of Δ. The responses are drawn from N(0,1) under the null hypothesis and from N(0,1) and N(Δ,1) under the alternative hypothesis. The nominal values of the significance level and power are $\alpha = 0.05$ and $1-\beta = 0.95$. In the left half of the table, the results are based on a test where $\kappa(f)$ is calculated for N(0,1). In the right half of the table $\kappa(f)$ is estimated throughout each trial. The mean and median value of D_n are shown. The exact value of $\gamma(f)$ is 0.141. The estimation starts at n_0=10 patients. Each result is based on $N = 10000$ simulations (except for $\Delta = 0.5$ when $N = 1000$).

Distribution	$\kappa(f)$ $\hat{\alpha}$	# pats mean median	$1-\hat{\beta}$	# pats mean median	K_n $\hat{\alpha}$	# pats mean median	$1-\hat{\beta}$	# pats mean median	D_n mean median
Normal (0,1)	0.0453	25.3 20.0	0.9503	25.5 20.0	0.0530	26.8 22.0	0.9492	27.0 22.0	0.141 0.134
95 % N(0,1)+ 5 % N(0,3)	0.0502	26.3 21.0	0.9446	26.6 21.0	0.0516	29.7 24.0	0.9448	29.2 24.0	0.134 0.127
90 % N(0,1)+ 10 % N(0,3)	0.0593	27.1 22.0	0.9356	27.4 22.0	0.0436	32.3 27.0	0.9453	32.5 26.0	0.127 0.121
80 % N(0,1)+ 20 % N(0,3)	0.0806	28.8 23.0	0.9164	29.2 24.0	0.0519	39.7 32.0	0.9484	39.9 32.0	0.114 0.109
95 % N(0,1)+ 5 % N(0,4)	0.0542	26.3 21.0	0.9419	26.7 22.0	0.0517	30.3 25.0	0.9486	30.0 24.0	0.133 0.126
Uniform $(0,\sqrt{12})$	0.0552	25.9 21.0	0.9431	26.0 21.0	0.0560	28.4 24.0	0.9440	28.0 23.0	0.133 0.129
Logistic $(b=\sqrt{3}/\pi)$	0.0366	24.4 20.0	0.9627	24.3 20.0	0.0468	24.5 20.0	0.9532	24.4 20.0	0.149 0.142
Cauchy $(r=0.54427)$	0.0458	25.7 21.0	0.9529	25.7 21.0	0.0238	31.1 26.0	0.9764	30.9 26.0	0.134 0.126
Double exp $(\lambda=\sqrt{2})$	0.0250	22.3 18.0	0.9765	22.4 18.0	0.0378	21.3 17.0	0.9660	21.2 17.0	0.164 0.154
Negative exp $(\lambda=1)$	0.0182	19.5 16.0	0.9827	19.5 16.0	0.0289	17.5 14.0	0.9735	17.7 15.0	0.185 0.172
Gamma $(a=1/\sqrt{2},\ d=2)$	0.0285	22.3 18.0	0.9717	22.1 18.0	0.0418	20.9 17.0	0.9600	21.0 17.0	0.163 0.155

Table 4

Simulated values of the significance level and power for different types of distributions. The nominal values are $\alpha = 0.05$ and $1-\beta = 0.95$. In the left half of the table $\kappa(f)$ is calculated under the normal distribution ($\kappa(f) = 1/(2\sqrt{\pi})$). In the right half of the table $\kappa(f)$ is estimated throughout the trial. The estimation starts at $n_0 = 10$ patients. In addition to the mean and median number of patients, the mean and median of D_n are shown. The alternative hypothesis is $\Delta = 1$. Each result is based on $N = 10000$ simulations.